DOCUMENT RESUME

ED 071 352 EM 010 497

TITLE	Introduction to Psychology and Leadership. Rank-Biserial Correlation as an Item Discrimination.
INSTITUTION	Naval Academy, Annapolis, Md.; Westinghouse Learning Corp., Annapolis, Md.
SPONS AGENCY	National Center for Educational Research and Development (DHEW/OE), Washington, D.C.
REPORT NO	TP-6-10
BUREAU NO	BR-8-0448
PUB DATE	11 May 70
CONTRACT	N00600-68-C-1525
NOTE	16p.; See also EM 010 418 and EM 010 419
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Comparative Statistics; *Correlation; Scores; *Statistical Analysis

ABSTRACT

Written as a technical report for the leadership course of the United States Naval Academy (see the final reports which summarize the course development project, EM 010 418, EM 010 419, and EM 010 484), this paper examines the use and interpretation of the rank-biserial correlation as an index of item discrimination. The advantages and disadvantages of this index are compared with those of alternative indices derived from the response of upper and lower groups divided on the basis of total test scores. Computational procedures and tests of statistical significance for the rank-biserial correlation are presented. Appropriate correction for the spurious correlation arising from the contribution of the item to total scores is also provided. (Author/SH)

Set #3

ED 071352

Westinghouse Learning Corporation

RANK-BISERIAL CORRELATION AS AN

ITEM DISCRIMINATION

IP 6.10 May 11, 1970

EM 010 497

ED 071352

TP-6.10

May 11, 1970

RANK-BISERIAL CORRELATION AS AN
INDEX OF ITEM DISCRIMINATION

Contract No. N00600-68-C-1525

Report No. TP

ABSTRACT

This paper examines the use and interpretation of the rank-biserial correlation as an index of item discrimination. The advantages and disadvantages of this index are compared with those of alternative indices derived from the response of upper and lower groups divided on the basis of total test scores.

Computational procedures and tests of statistical significance for the rank-biserial correlation are presented. Appropriate correction for the spurious correlation arising from the contribution of the item to total scores is also provided.

Prepared by:

David W. Bessemer

Approved by:

Project Manager
Leadership Course

WESTINGHOUSE LEARNING CORPORATION
2083 WEST STREET
ANNAPOLIS, MARYLAND 21401

Brennan (1969) has discussed the use and interpretation of item discrimination indices in the evaluation of criterion-referenced tests. He recommends the use of the index D, and a more general index B_i, both of which represent the difference in percent correct between upper and lower groups of students dichotomized on the basis of total test score. In the case of D, the students are divided into two equal groups, while B_i permits the use of any two group sizes.[1] Brennan rightly points out three important features which recommend the use of D and B_i: (1) they measure degree of discrimination in direct correspondence to a widely acceptable intuitive notion of the meaning of discrimination, (2) they are easily computable and interpretable by unsophisticated users, and (3) they are distribution free, and do not require questionable assumptions or hazardous approximations in their tests of significance.

There are, however, three aspects of D and B_i which seriously detract from their value as measures of item discrimination. First, the dichotomization of the total score variable discards information on discriminations among students in the upper group and among students in the lower group. This results in indices largely sensitive to discrimination in the region of the division between upper and lower groups. As Brennan himself points out, groups used in the evaluation of criterion-referenced tests are rarely large, so that any substantial loss of

---

[1] At times Brennan appears to confuse defects of proposed tests of significance for D with defects of D as a measure of discrimination. Since D is only a special case of B_i, any advantage claimed for B_i is equally true of D, and any test procedure recommended for B_i is equally applicable to D.

information is strictly to be avoided. Secondly, the use of D and B_i requires the evaluator to select a cutoff between lower and upper groups. No criteria for this selection have been offered, so that even the most experienced evaluator is confronted with a serious problem of judgment. Furthermore, since the values of the B_i indices are markedly affected by the cutoff decision, the comparability of the B_i indices from one test to another is impaired.

Finally, there is a third difficulty which is not unique to D and B_i, but which is shared by most indices of item discrimination. That difficulty is the spuriously high correlations which result from the fact that the item itself contributes to the total score. Unless a correction is introduced, obtained values of D and B_i are positively biased, and the bias may be pronounced when only a few items contribute to total scores, as is usually the case for short criterion-referenced tests.

A discrimination index based on rank order correlation will be presented in the sections which follow. It will be shown to retain the advantages of the D and B_i indices, while avoiding their defects.

Rank-biserial correlation

A measure of correlation between a ranked variable and a dichotomy was developed by Cureton (1956, 1968). This measure, called the rank-biserial correlation, r_{rb}, is functionally analogous to the point-biserial r, but is closely related to Kendall's tau, being based on the number of agreements and disagreements in rank order between the two variables. For the purpose of determining correspondence between rank orders, the dichotomy is considered to be a categorization into two ranks with multiple ties (Whitfield, 1947).

Consider the tabulation given below, where Y is the rank variable (for example, ranks of the students on total score) and X is the dichotomy ($X = 1$ representing a correct response to a test item, $X = 0$ an incorrect response). X and Y agree in ranking any pair of students when the higher ranked on Y obtained a correct response, and the lower ranked

Y (Ranks)

$X = 0$			3		5	6		8	9	10
$X = 1$	1	2		4			7			

obtained an incorrect response. Thus for every rank with $X = 1$, there is an agreement for every <u>lower</u> rank which appears with $X = 0$. The number of agreements for the ranks of 1, 2, 4, and 7 are 6, 6, 5, and 3 respectively.

On the other hand, a disagreement in rank occurs when the student higher ranked on Y obtained an incorrect response, and the lower ranked obtained a correct response. Thus, for every rank with $X = 0$, there is a disagreement for every <u>lower</u> rank with $X = 1$. The number of disagreements for the ranks of 3, 5, 6, 8, 9, and 10 are 2, 1, 1, 0, 0, and 0 respectively.

Cureton defined r_{rb} as follows: $r_{rb} = (P-Q)/P_{max}$ where P is the total number of agreements, Q is the total number of disagreements, and P_{max} is the maximum possible value of P. It should be noted that the numerator or r_{rb} is the same as the numerator of Kendall's tau, the two measures differing only in the denominator. In the case that no ties on Y occur $P_{max} = n_1 n_0$, where n_1 is the number of ranks having $X = 1$, and n_0 is the number of ranks having $X = 0$. The denominator of r_{rb} was chosen to insure that the possible range of values of 1 was obtainable under all circumstances. Kendall's tau does not necessarily attain these limits when the relationship between X and Y is perfect. For example, when $n = 5$, $n_1 = 2$, and $n_2 = 3$, and the ranks 1 and 2 have scores of $X = 1$, $r_{rb} = 1$, but $\tau = .77$. For the data above, $r_{rb} = \frac{20-4}{(4)(6)} = .67$.

Glass (1965, 1966) presented a simplified computational procedure for r_{rb} useful when no ties on Y occur. Rather than counting agreements and disagreements as above, it is necessary only to compute $\bar{Y}_1$ and $\bar{Y}_0$, the mean ranks for $X = 1$ and $X = 0$ respectively. Then $r_{rb} = 2(\bar{Y}_0 - \bar{Y}_1)/n$. However, this simplification will rarely apply to short criterion-referenced tests, particularly when the sample size is large, since many students will obtain identical total scores and be assigned tied ranks.

Computation of r_{rb} with tied ranks

A correction to P_{max} must be made when ranks are tied. If there is a perfect relationship between X and Y, agreements are lost among tied ranks at the point of division between the n_1 upper ranks and the n_0 lower ranks. In the example given below, there are 6 ranks tied at the point of division, four having $X = 1$ and 2 having $X = 0$.

X = 0					4.5	4.5	8	9	10
X = 1	1	4.5	4.5	4.5	4.5				

There are (4)(2) = 8 possible agreements lost as a result of the tied ranks. If t_1 is the number of ranks with $X = 1$ tied at the division, and t_0 the number with $X = 0$, then $P_{max} = n_1 n_0 - t_1 t_0$.

Computation of r_{rb} in a frequency distribution

When a bivariate frequency distribution is available, a simple computational procedure may be followed which incorporates the correction for ties, and even avoids the assignment of ranks to the Y variable.

The notation for the bivariate distribution shown below represents the frequency of correct and incorrect responses for each possible total score, along with cumulative frequencies for correct and incorrect responses.

Total Score	Correct Frequency	Cumulative Frequency	Incorrect Frequency	Cumulative Frequency
Y_k	$f_{1,k}$	$F_{1,k}$	$f_{0,k}$	$F_{0,k}$
Y_{k-1}	$f_{1,k-1}$	$F_{1,k-1}$	$f_{0,k-1}$	$F_{0,k-1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y_i	$f_{1,i}$	$F_{1,i}$	$f_{0,i}$	$F_{0,i-1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y_2	$f_{1,2}$	$F_{1,2}$	$f_{0,2}$	$F_{0,2}$
Y_1	$f_{1,1}$	$F_{1,1}$	$f_{0,1}$	$F_{0,1}$

Note, of course, that the cumulative frequencies $F_{1,i} = \sum_{j=1}^{i} f_{1,j}$ and $F_{0,i} = \sum_{j=1}^{i} f_{0,j}$. We will also require a symbol F_i for the marginal cumulative frequency, $F_i = F_{1,i} + F_{0,i}$. Then the number of agreements are $P = \sum_{i=1}^{k} f_{1,i} F_{0,i-i}$, and the number of disagreements $Q = \sum_{i=1}^{k} f_{0,i} F_{1,i-1}$.

To obtain t_0 and t_1, examine the marginal frequencies to find F_i^* and F_{i-1}^* for which F_i^* n_0 F_{i-1}^*. Then $t_1 = n_1 - F_i^*$ and $t_0 = n_0 - F_{i-1}^*$.

Since $n_1 = F_{1,k}$ and $n_0 = F_{0k}$, $P_{max} = n_1 n_0 - t_1 t_0 = F_{1,k}F_{0,k} - (F_{1,k} - F_i^*)(F_{0,k} - F_{i-1}^*) = F_{1,k}F_{0,k} - F_{1,k}F_{0,k} + F_{1,k}F_{i-1}^* + F_{0,k}F_i^* - F_i^*F_{i-1}^* = F_{1,k}F_{i-1}^* + F_{0,k}F_i^* - F_i^*F_{i-1}^*$. Thus r_{rb} becomes, in frequency distribution notation,

$$r_{rb} = \frac{\sum_{i=1}^{k} (f_{1,i}F_{0,i-1} - f_{0,i}F_{1,i-1})}{F_{1,k}F_{i-1}^* + F_{0,k}F_i^* - F_i^*F_{i-1}^*}$$

In the example distribution given below $P = 73$, $Q = 20$ $F_i^* = 12$, $F_{i-1}^* = 8$, $F_{1,k} = 12$, and $F_{0,k} = 8$.

$$\text{Thus } r_{rb} = \frac{73-20}{(12)(8) + 8(12) - (8)(12)} = \frac{53}{96} = .55$$

Y_i	$f_{1,i}$	$f_{0,i}$	$F_{1,i}$	$F_{0,i}$	F_i	$f_{1,i}F_{0,i-1}$	$F_{0,i}F_{1,i-1}$
10	1	0	12	8	20	8	0
9	1	0	11	8	19	8	0
8	1	0	10	8	18	8	0
7	1	1	9	8	17	7	8
6	2	1	8	7	15	12	6
5	4	0	6	6	12	24	0
4	0	3	2	6	9	0	6
3	2	0	2	3	5	6	0
2	0	1	0	3	3	0	0
1	0	1	0	2	2	0	0
0	0	1	0	1	1	0	0
						73	20

This value may be compared with $r_{pb} = .50$ and $D = .40$ for the same data. Also $B_1 = .63$, $B_2 = .67$, $B_3 = .71$, $B_4 = .67$, $B_5 = .58$, $B_6 = .25$, $B_7 = .27$, $B_8 = .47$, $B_9 = .44$ and $B_{10} = .42$ where the subscript refers to the lowest value of Y_i included in the "upper" group. It is interesting to note that the highest values of B_i occur with cutoffs below the median, whereas most evaluators would place the cutoff above the median in distinguishing "acceptable" from "unacceptable" levels of performance.

Correction for spurious correlation

Like other item discrimination indices, r_{rb} will be subject to spurious correlation arising from the contribution of the item to the total score, if the computational procedure given above is followed.

However, the formula for the frequency distribution computation is easily modified to eliminate the bias due to suprious correlation. Since the total score is increased by one for those who have a correct response on the item, the same computation procedure may be followed if the total score is simply reduced by one for all students having a correct response. In terms of the frequency distribution, the reduction simply requires each frequency and cumulative frequencies in the columns for correct response to be shifted down to the next lower score, and the computation of a new set of marginal cumulative frequencies. If this is done, the formulas (still using the original notation) become:

$$P = \sum_{i=1}^{k} f_{1,i}F_{0,i-2} \text{ and } Q = \sum_{i=1}^{k} f_{0,i}F_{1,i}$$

$$P_{max} = F_{1,k}(F^{*}_{1,i-2} + F^{*}_{0,i-1}) + F_{0,k}(F^{*}_{1,i-1} + F^{*}_{0,i})$$

$$- (F^{*}_{1,i-2} + F^{*}_{0,i-1})(F^{*}_{1,i-1} + F^{*}_{0,i})$$

$$\text{where } F^{*}_{1,i} + F^{*}_{0,i-1} \geq .50 \geq F^{*}_{1,i-1} + F^{*}_{0,i}$$

For the example above

$P = 8 + 8 + 7 + 6 + 12 + 12 + 4 = 57.$

$Q = 9 + 8 + 6 = 23$

$F^{*}_{1,i-1} + F^{*}_{0,i} = 6 + 6 = 12$

$F^{*}_{1,i-2} + F^{*}_{0,i-1} = 2 + 3 = 5$

Then $$r_{rb} = \frac{57 - 23}{(12)(5) + (8)(12) - (12)(5)} = \frac{34}{96} = .35,$$

in comparison with the uncorrected value $r_{rb} = .55$.

The example demonstrates that the effect of spurious correlation can be very substantial when the number of items is small. It is recommended that the corrected formula

$$r_{rb} = \frac{\sum_{i=1}^{k} (f_{1,i-1}F_{0,i-2} - f_{0,i}F_{1,i})}{F_{1,k}(F^*_{1,i-2} + F^*_{0,i-1}) + F_{0,k}(F^*_{1,i-1} + F^*_{0,i}) - (F^*_{1,i-2} + F^*_{0,i-1})(F^*_{1,i-1} + F^*_{0,i})}$$

to be used whenever r_{rb} is used as an item discrimination index.

Tests of significance

Several different approaches may be taken in testing the statistical significance of r_{rb}. Cureton (1956) suggested that the Mann-Whitney U-test be used for this purpose (see Siegel, 1956). The value of U employed as a test-statistic corresponds to the smaller of the values of P or Q as computed above. The tables of critical values of U given in Siegel's book provide exact tests so long as $n_0 \leq 20$ and $n_1 \leq 20$, and no ties appear in the ranked variable.

In the case of ties, when $n_0 \leq 8$ or $n_1 \leq 8$, an appropriate procedure is to perform an exact randomization test on P. The value of P is determined for each of the $\binom{n}{n_1} = \frac{n!}{n_1!n_0!}$ randomizations of the ranks between the values of the dichotomy, with the restriction that n_1 ranks are assigned to X = 1 and n_0 ranks to X = 0. For an α % test, the distribution of possible values of P is used to determine if $\frac{1}{2}\alpha$ % or less of the values are equal to or more extreme than the observed value of P. If this is the case, the observed value is declared significant.

Except for very small values of $n = n_0 + n_1$, or extreme splits between n_0 and n_1, the computational labor of the exact randomization test is excessive due to the large number of values of P to be computed, even when performed by a digital computer. Where the cost of computer time is excessive, the only alternative available is the approximate "jackknife" technique. The details of a "jackknife" solution are too extensive to be presented here, and the reader is referred to the discussion by Mosteller and Tukey (1968).

When $n_1 > 8$ and $n_2 > 8$, whether or not ties are present, a very satisfactory normal approximation may be employed. Under the null hypothesis, P - Q will be approximately normally distributed with mean $\mu = 0$ and variance

$$\sigma^2_{P-Q} = \frac{n_1 n_0}{3n(n-1)} \left[n^3 - n - \sum_{i=1}^{k} (f_i^3 - f_i) \right]$$

as given by Kendall (1962), where f_i refers to the marginal frequency of occurrence of Y_i. The approximation is further improved by the incorporation of a correction for continuity reducing P - Q in absolute value. Thus the test statistic $F = \frac{|P - Q| - C}{\sigma_{P-Q}}$ may be referred to tables of the unit normal distribution, where C is the value of the correction for continuity.

When no tied ranks are present C = 1. In other cases an approximate correction suggested by Kendall (1962) may be obtained from the following formula. Let Y_h and Y_1 be the highest and lowest scores in the distribution with $f_h > 0$ and $f_1 > 0$, respectively. Then

$C = \frac{2n - f_h - f_1}{2(g - 1)}$ where g is the number of distinct Y_i with $f_i > 0$.

The value of C given by this formula is one-half of the <u>average</u> distance between adjacent possible values of P - Q.

In the example above, the value of P - Q = 34, when corrected for spurious correlation. Values of f_1 through f_{10} are 1, 1, 3, 0, 7, 2, 2, 2, 1, and 1, respectively, using $f_i = f_{1,i-1} + f_{0,i}$. Then

$$\sigma^2_{P-Q} = \frac{(12)(8)}{(3)(20)19}\left[20^3 - 20 - (3^3 - 3) - (7^3 - 7) - 3(2^3 - 2)\right]$$

$$= (8/95)\left[7980 - 24 - 336 - 3(6)\right]$$

$$= 8(7602)/95 = 640.168$$

and $\sigma_{P-Q} = 25.30$. For the highest and lowest score $f_h = f_l = 1$, and the number of di.cinct scores occurring is eight, giving $g - 1 = 7$. Then

$$C = \frac{2(20) - 1 - 1}{2(7)} = \frac{19}{7} = 2.71$$

and $Z = \frac{34 - 2.71}{25.30} = \frac{31.29}{25.30} = 1.24$ indicating that r_{rb} is not significant at $\alpha = .05$. It should be noted that the correction for continuity is quite important in applications of r_{rb} to tests with only a few items, as illustrated here.

<u>Comparison of r_{rb} with D and B_i</u>

The basic nature of r_{rb} is quite similar to D and B_i in several respects. All are based on the same intuitive notion of discrimination, i.e., that an item discriminates (positively) between individuals whenever their difference in response to the item corresponds to their difference in performance as based on total score. The value of r_{rb} is subject to the following simple interpretation: r_{rb} is an estimate of the difference between the probability that the rank order of two randomly selected individuals on total score and item will be in agreement, and the probability that their rank order on total score and item will be in disagreement. The D and B_i indices are subject to exactly the same interpretation, except that only two ranks of performance are recognized on the basis of total score, i.e., an upper level and a lower level. In fact, the computational formulas for D and B_i are merely special cases of the r_{rb} formula when the rank ordering is dichotomized. Since r_{rb}, D, and B_i are based only on ordering of performance, not on arithmetic distances between performance levels, all are entirely distribution-free, being invariant under any monotonic transformation of the total scores.

D and B_i are slightly easier to compute, particularly when ties are present in the rank order on total score. However, this computational simplicity is purchased at the expense of information lost as a result of the dichotomization of the total score ranking. There does not seem to be any logical reason that an index of discrimination should ignore discriminations among students in the upper group, and among students in the lower group. Since r_{rb} incorporates all possible information on discrimination obtainable from a rank ordering, it is to be preferred on that basis if no other. Furthermore, r_{rb} avoids entirely the difficulty of judging an appropriate point of dichotomization of the total scores which is involved in D and B_i.

The remaining advantages of r_{rb} concern technical statistical properties. The dichotomization involved in D and B_i produce indices with greater sampling variability and tests of significance of lesser power-efficiency. The power of the Mann-Whitney U used to test r_{rb} is approximately 95% against normal alternatives.

The power-efficiency of the median test, which corresponds to the test for D, is about 95% for $n = 6$, declining to an asymptotic value of 63% as n increases. Thus a sample size considerably larger than that used with r_{rb} is required if the test of D is to have equivalent power, unless the sample sizes are very small.

Finally, D and B_i, as presented by Brennan (1969), have not been modified to correct for spurious correlation. This fact not only produces a positive bias in the reported values of the indices, but also invalidates the test of significance presented by Brennan. While it would be no more difficult to modify the computation of D and B_i and their test procedure than it was to modify r_{rb} and its test, the general superiority of r_{rb} would seem to make unnecessary the additional effort required to develop such modifications.

References

Brennan, R. L. A discrimination index for criterion-referenced tests. Annapolis, Md.: United States Naval Academy Multi-Media Course Development Memo 2.13. December 1, 1969.

Cureton, E. E. Rank-biserial correlation. Psychometrika, 1956, 21, 287-290.

Cureton, E. E. Rank-biserial correlation when ties are present. Educational and Psychological Measurement, 1968, 28, 77-79.

Glass, G. V. A ranking variable analogue of biserial correlation: Implications for shortcut item analysis. Journal of Educational Measurement, 1965, 2, 91-96.

Glass, G. V. Note on rank-biserial correlation. Educational and Psychological Measurement, 1966, 26, 623-631.

Kendall, M. G. Rank correlation methods. London: Griffin, 1962 (3rd. ed.)

Seigel, S. Nonparametric statistics for the behavioral sciences. New York: McGraw-Hill, 1956.

Whitfield, J. W. Rank correlation between two variables, one of which is ranked, the other dichotomous. Biometrika, 1947, 34, 292-296.